Methods of Electromagnetic Field Analysis*

By S. A. SCHELKUNOFF

This paper presents a discussion of ideas involved in various mathematical methods of electromagnetic field analysis and of the inter-relations between these ideas. It stresses the points of contact between circuit and field theories and their mutually complementary character. While the field theory focuses our attention on the electromagnetic state as a function of position in space, the generalized circuit theory is preoccupied with the electromagnetic state as a function of time. The points of contact between the field and circuit theories are many. Thus, Maxwell's equations are identical with Kirchhoff's equations (really Lagrange-Maxwell equations) of certain three-dimensional networks in which only the adjacent meshes are coupled. The integral equations for the electrical current in conductors embedded in dielectric media are also Kirchhoff equations of certain networks containing infinitely many meshes with a coupling between every two meshes.

From the point of view of electrical performance the difference between a physical network of lumped elements and a continuous network, such as a resonator, is due to a certain difference in the distribution of the zeros and poles of associated impedance functions in the complex impedance plane. Similarly, the difference between ordinary transmission lines and wave guides is due to a difference in the distribution of natural propagation constants.

The paper ends with a general discussion of the discontinuities in wave guides, idealized boundary conditions for simplification of electromagnetic problems, and the analytical character of field vectors regarded as functions of the complex oscillation constant.

IN THE last few years engineering applications of electromagnetic field theory have been greatly expanded. Field theory has become essential for the solution of many practical problems and in planning engineering experiments. New applications have influenced the theory itself and have led to new conceptions. The chasm between the circuit theory of low frequency electrical phenomena and the field theory of high-frequency phenomena has disappeared. The two theories have met in wave guides and their merger has become essential. This paper is a discussion of the essential ideas underlying various mathematical methods of analysis of electromagnetic oscillations and waves in the light of new applications and of the merger of the originally distinct circuit and field theories.

Circuit Theory

Circuit theory is a mathematical method and it should not be confused with circuits. Empty space is neither a circuit nor a network; but as we shall soon see, for the purposes of analysis the empty space can be treated as a network. It is perfectly true that until recently circuit theory was con-

* This paper was originally delivered as a lecture at a meeting sponsored by the Basic Science Group of the American Institute of Electrical Engineers, April 12, 1945.

cerned almost exclusively with aggregates of "circuit elements" interconnected in various ways. It is also true that the most familiar form of circuit equations is that which is similar to Kirchhoff's equations for the steady current flow in networks of conducting rods, published[1] in April 1845.

This form is applicable only to circuits. However, the application of these "Kirchhoff equations" to alternating currents, natural as it may seem to us now, was not obvious one hundred years ago. The first equation for a simple circuit consisting of a capacitor, an inductor, and a resistor in series was published in 1853 by Lord Kelvin.[2] Interestingly enough his approach is based on the ideas applicable both to conventional circuits and to high-frequency resonators. If q is the electric charge on one plate of the capacitor, the energy stored in the capacitor is $q^2/2C$, where the coefficient C depends on the geometry of the capacitor. The magnetic energy of the circuit is $\frac{1}{2} L\dot{q}^2$, where $\dot{q}$ is the time rate of change of the charge, that is, the current in the circuit, and L is a coefficient depending on the geometry of the circuit. The rate of energy transformation into heat is $R\dot{q}^2$, where R is a coefficient depending on the geometry of the conductors (and of course on their resistivity). The law of conservation of energy demands that

$$\frac{d}{dt}\left[q^2/2C + \tfrac{1}{2}L\dot{q}^2\right] = -R\dot{q}^2. \tag{1}$$

When the differentiation is performed and $\dot{q}$ is cancelled, the usual form of the equation is obtained. The coefficients of proportionality, that is, the inductance L, the capacitance C, and the resistance R sum up and stress the really important electrical characteristics of the circuit; the details of the construction of the circuit are suppressed.

It was Maxwell who formulated the general equations for electric networks by extending the application of a method developed by Lagrange for mechanical systems. This Maxwell did in his last two lectures. In the words of his student, J. H. Fleming:[3] "Maxwell, by a process of extraordinary ingenuity, extended this reasoning (the method of Lagrange) from materio-motive forces, masses, velocities and kinetic energies of gross matter to the electromotive forces, quantities, currents, and electrokinetic energies of electrical matter, and in so doing obtained a similar equation of great generality for attacking electrical problems."

Before discussing the Lagrange-Maxwell method more completely, let us see if we can construct a network whose electrical properties would be the same as those of a continuous medium.

[1] *Annalen der Physik.*
[2] *Philosophical Magazine.*
[3] *Philosophical Magazine*, 1885.

Natural Network Models of Continuous Media and Maxwell's Differential Equations

Transmission line theory represents a well known example of the application of circuit theory to continuous systems. Two-wire transmission lines are subdivided into infinitesimal sections by planes perpendicular to the lines. Each section is replaced by a capacitor whose capacitance is so chosen that, for a given voltage across the transmission line, the electric charges on the plates of the capacitor are correspondingly equal to the charges on the sections of the wires constituting the line. The leads connecting the terminals of these capacitors are then assumed to possess an inductance and a resistance but no capacitance. Thus the electric flux or displacement is "swept" into tiny capacitors, and the magnetic flux or displacement into tiny inductors.

This representation is good only at low frequencies because it depends on the assumption that the electric displacement is only in one direction, namely at right angles to the transmission line. In effect, this representation neglects the capacitance between different parts of the same conductor and includes only the capacitance between the opposite segments of different conductors. That is, while we have recognized that the inductance and capacitance are distributed in the direction parallel to the transmission line, we have ignored the fact that they are also distributed at right angles to the line. In the general representation we should subdivide the medium into infinitesimal blocks and devise a three-dimensional network lattice of infinitely small meshes, Fig. 1. The displacement current can be swept equally into tiny capacitors. If the medium is dissipative, the resistors may be inserted in parallel with the capacitors to take care of the conduction currents in the medium. The magnetic flux is swept equally into tiny coils in the corners of each mesh. However, the resulting network is not homogeneous. Besides meshes of type A consisting of four capacitors and four inductors, it contains meshes of type B consisting of inductors only; and yet we started with a homogeneous medium. Gabriel Kron solved the difficulty by introducing ideal transformers (with one-to-one turn ratio) with their windings in series with the coils at the opposite corners of each A-mesh. These transformers do not affect the electrical performance of the A-meshes but introduce infinite impedance into B-meshes and thus effectively eliminate them.

As a matter of fact, such transformers should properly be included in the network representations of two-wire lines. In fact, by implication they *are* included as soon as we state that the direct and return currents in the line are equal and opposite. Without an infinite impedance to currents flowing in the same direction we cannot have the balance. Pursuing the matter

further, we should say that all this is in accord with physical facts. The inductance per unit length of an *infinitely long* isolated wire is infinite. The mutual inductance between two parallel wires is also infinite. The two wires are the "windings" of an ideal transformer and a finite impedance is presented only to equal and opposite currents. In the case of wires of finite length the essentially three-dimensional character of the structure manifests itself, and other modes of propagation have to be considered.

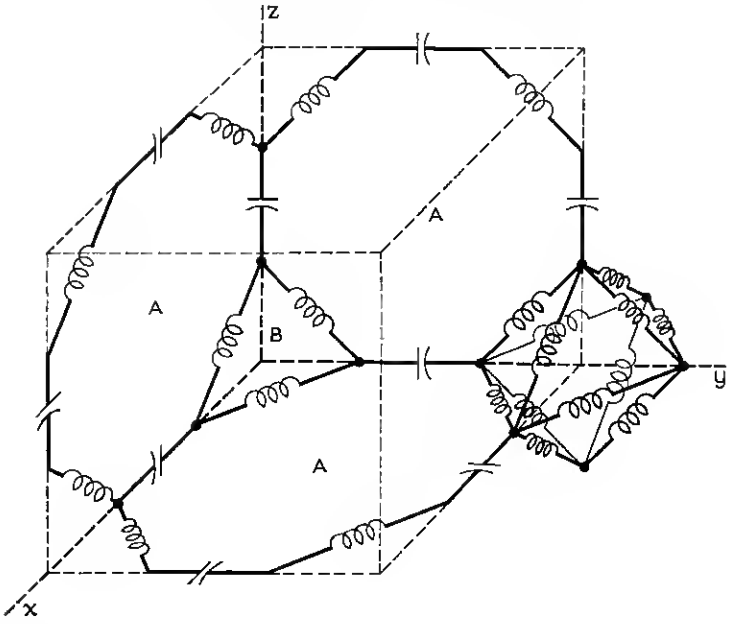

Fig. 1—Typical equivalent meshes in a circuit representation of continuous media.

It is evident that the homogeneity of the medium is not a prerequisite for the existence of its network model. Having the values of L and C at our disposal, we can choose them to reflect the dependence of the permeability μ and the dielectric constant ϵ on position.

If we divide the medium into small blocks of volume $\Delta x\,\Delta y\,\Delta z$, the capacitance C_x of the typical capacitor in those branches of the network which are parallel to the x-axis is $C_x = \epsilon\,\Delta y\,\Delta z/\Delta x$, where ϵ is the dielectric constant. The conductance in parallel with this is $G_x = g\Delta y\Delta z/\Delta x$. The inductance of the typical coil in the xy-plane is $L_{xy} = \mu\,\Delta x\Delta y/4\Delta z$. The voltages across the capacitors are $E_x\Delta x$, $E_y\Delta y$, $E_z\Delta z$, where E_x, E_y, E_z are the electric intensities, that is, the voltages per unit length in the respective directions. The currents in the coils situated in the xy-plane are equal to $H_z\Delta z$; simi-

larly the currents in the other coils are $H_x \Delta x$ and $H_y \Delta y$. It is to be noted that the capacitors are associated with the corresponding longitudinal components of the electric field while the inductors go with the transverse components of the magnetic field. Applying Kirchhoff's laws to the network in Fig. 1, we should and do obtain Maxwell's field equations. Similarly, we can construct network lattices in the patterns of other coordinate systems, cylindrical and spherical, for example.

Among the obvious conclusions to be drawn from this analysis of the network structure of the medium supporting the electromagnetic field is the validity of certain general network theorems such as the Reciprocity Theorem and Thevenin's Theorem.

Reduced Network Models and Integral Equations of Lorentz Type

So far we have been concerned with the electromagnetic field in its entirety. In order to visualize the medium as a three-dimensional network we have selected the most direct course: We have subdivided the medium into blocks of displacement current, compressed them into capacitors, and eliminated displacement currents from the rest of space; similarly, we have

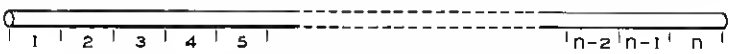

Fig. 2—Subdivision of a straight antenna for its representation by a reduced network with n meshes.

swept the magnetic flux into neat little packages. But this is not the only course open to us. We can suppress the medium just as completely as we normally do in the analysis of elementary networks. In order to illustrate this method let us consider a doublet antenna, Fig. 2. We shall divide it into n sections. The current and charge in any one section exert forces on the charge in any other section. We can regard each section of the antenna as a mesh of a network in which every mesh is coupled to every other mesh. In each mesh the voltage which is necessary to compensate for the electromotive force of self-induction of the mesh itself, for the resistance of the mesh (or rather for the internal impedance of the wire), and for the voltages induced from all the other meshes, is the impressed voltage. The equations assume the following form:

$$\begin{aligned} Z_{11}I_1 + Z_{12}I_2 + Z_{13}I_3 + \cdots + Z_{1n}I_n &= V_1, \\ Z_{21}I_1 + Z_{22}I_2 + Z_{23}I_3 + \cdots + Z_{2n}I_n &= V_2, \\ Z_{n1}I_1 + Z_{n2}I_2 + Z_{n3}I_3 + \cdots + Z_{nn}I_n &= V_n, \end{aligned} \tag{2}$$

where the I's are the currents in the various sections of the antenna and the V's are the *impressed* voltages. The Z's are the self-impedances and the mutual impedances, and are calculated from the law of force between two charged particles. In a transmitting antenna the impressed voltage is zero everywhere except in a restricted region. In the receiving antenna the voltage is impressed on all sections; but one section, the "load," has a very different self-impedance from the remaining sections.

When n is finite, our equations are approximate. If we make n infinite and introduce the impressed electric intensity, that is, the impressed voltage per unit length, we convert equations (2) into a single integral equation. More generally we may have to consider the transverse dimensions of the antenna and divide the entire surface of the antenna into elementary surface elements, each of which will represent *two* meshes in our network. We have to have two meshes for each surface element because the current may in general change its direction from point to point and in order to specify it completely we must consider two components of the current. These may be taken as tangential to some Gaussian coordinate lines drawn on the surface of the antenna. The exact network equations will appear as a system of two integral equations involving double integrals.

In this discussion, we have assumed that the medium outside the antenna is homogeneous. No difficulty is presented by the simultaneous inclusion of a transmitting and a receiving antenna. The two form just one network and the voltages impressed on the various meshes of the receiving antenna represent simply the coupling between these meshes and the meshes of the transmitting antenna. All the mutual impedances are calculable from the general equation,

$$E = -\mu \frac{\partial A}{\partial t} - \text{grad } V, \tag{3}$$

representing the force per unit charge due to a given moving charge. If we so desire, we can take equation (3) with the explicit expressions for A and V in terms of electric current and charge as the fundamental equations of electromagnetic theory and dispense with Maxwell's differential equations altogether. This course is feasible but inexpedient. Actual applications of this equation turn out to be much too complicated in the great majority of practical problems. It is only when we already know the current and charge distribution that (3) becomes really useful. Thus in the accepted development of electromagnetic theory (3) is subordinated to Maxwell's equations and derived from them.

Normalized Network Model and Lagrange-Maxwell Electrodynamical Equations

Let us now return to the ideas of Lagrange as applied to electromagnetics. In dynamics the Lagrange equations are formulated in terms of the kinetic energy T expressed as a function of velocities, potential energy U expressed as a function of coordinates, and a dissipation function F expressed as a function of velocities. In network theory T is the magnetic energy expressed as a function of currents, U is the electric energy expressed in terms of charges, and F is the dissipation function in terms of currents. Lagrange-Maxwell equations are then written in the following form

$$\frac{d}{dt}\left[\frac{\partial}{\partial \dot{I}_n}(T - U)\right] - \frac{\partial}{\partial q_n}(T - U) + \frac{\partial F}{\partial \dot{I}_n} = V_n, \tag{4}$$

where I_n is the typical mesh current, q_n is its time integral, and V_n is the *impressed* electromotive force, that is, the electromotive force not accounted for by the magnetic induction and the charges in the network. The various functions in the equation are

$$T = \Sigma_m \Sigma_n \tfrac{1}{2} L_{mn} I_m I_n, \qquad U = \Sigma_m \Sigma_n \frac{q_m q_n}{2C_{mn}}, \qquad F = \Sigma_m \Sigma_n \tfrac{1}{2} R_{mn} I_m I_n, \tag{5}$$

where L_{mn} is the mutual inductance between two typical meshes (the self-inductance if $m = n$), C_{mn} is the mutual capacitance and R_{mn} is the mutual resistance. The mesh currents are introduced in order to insure that the total current either entering or leaving a typical junction of the network elements is zero. If we perform the differentiations indicated in equation (4), we shall obtain the network equations in their usual form.

Let us now suppose that $F = 0$ and $V_n = 0$. In higher algebra it is shown that by a linear transformation two quadratic functions, T and U for example, can be reduced to normal forms in which there are no mutual terms

$$T = \Sigma_n \tfrac{1}{2} L_n \hat{I}_n^2, \qquad U = \Sigma_n \hat{q}_n^2 / 2C_n. \tag{6}$$

In this case equations (4) will assume the following simple form

$$L_n \frac{d\hat{I}_n}{dt} + \frac{\hat{q}_n}{C_n} = 0. \tag{7}$$

It is as if we had a certain number of isolated single-mesh circuits. Equations (7) represent the *normal modes of oscillation* of the network.

Take the simple case of two identical coupled circuits, Fig. 3. The network equations are

$$L\frac{d^2I_1}{dt^2} + \frac{I_1}{C} - M\frac{d^2I_2}{dt^2} = 0, \qquad -M\frac{d^2I_1}{dt^2} + \frac{I_2}{C} + L\frac{d^2I_2}{dt^2} = 0. \quad (8)$$

It is evident by inspection that there are two possible modes of oscillation. In one mode $I_1 = I_2$ and in the other $I_1 = -I_2$. The natural frequency of the first mode is $\omega_1 = 1/\sqrt{(L-M)C}$ and that of the second mode $\omega_2 = 1/\sqrt{(L+M)C}$. The magnetic energy function is

$$\begin{aligned} T &= \tfrac{1}{2}LI_1^2 - MI_1I_2 + \tfrac{1}{2}LI_2^2 \\ &= \tfrac{1}{2}(L-M)\left[\frac{I_1+I_2}{\sqrt{2}}\right]^2 + \tfrac{1}{2}(L+M)\left[\frac{I_1-I_2}{\sqrt{2}}\right]^2. \end{aligned} \quad (9)$$

Thus the sum and the difference of the currents in the two meshes oscillate independently.

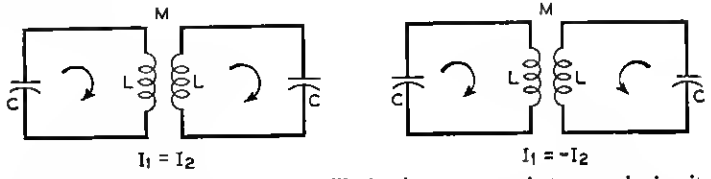

Fig. 3—Two possible modes of oscillation in a symmetric two-mesh circuit.

More generally a network with n meshes possesses n independent modes of oscillation. In eacb mode the ratios of the mesh currents $I_1, I_2, \cdots I_n$ are prescribed by the network parameters and the connections of the network elements, but the relative strength of the oscillation remains arbitrary. When we pass to networks with distributed parameters such as sections of transmission lines and cavity resonators, we find merely that the number of independent modes of oscillation is infinite. In the case of a nondissipative uniform transmission line with both ends shorted, the natural frequencies of the various oscillation modes are proportional to the sequence of integers: 1, 2, 3, ...The current distribution for the n-th mode is given by $\sin(n\pi x/\ell)$, where ℓ is the length of the section; but the actual amplitude remains arbitrary. For the gravest mode ($n = 1$) the middle part of the line section behaves as a capacitor and the ends as inductors. For the higher modes the line is subdivided into sections, some of which act primarily as capacitors and others as inductors.

In the case of cavity resonators of some simple shapes, such as parallelopipedal, cylindrical and spherical, the determination of the oscillation modes is a fairly simple prohlem. The dynamical equations of the resonator

(Maxwell's field equations) are partial differential equations. Their solutions would normally involve arbitrary functions; but since the tangential electric intensity vanishes at the conducting boundary of the resonator, the solutions assume a much less arbitrary form involving only an infinite set of arbitrary constants. Particular solutions are sought in the form of products of three functions, each depending on only one coordinate. For parallelopipedal cavity resonators the various components of electric and magnetic intensity are assumed in the form $X(x)\ Y(y)\ Z(z)$. By substituting in Maxwell's equations it is found—very fortunately indeed—that X, Y, Z may be obtained as solutions of ordinary differential equations. The boundary conditions at the boundaries of the box $x = 0,a; y = 0,b; z = o,c$ are easy to satisfy because we have to work with only one of these three functions at a time.

In general, however, the problem of calculating oscillation modes is by no means simple; but once these modes have been determined, the problem of forced oscillations as well as free oscillations is practically solved. For instance, a small loop inside a resonator is coupled to the various modes and the coupling coefficients can be determined by evaluating the flux linkages.

Every physical circuit possesses an infinite number of degrees of freedom and circuits with a finite number of degrees of freedom are abstractions. If we take special measures to concentrate magnetic energy as much as possible in a few regions of the medium and electric energy in a few other regions, we shall have a physical network in which a finite number of oscillation modes will be well separated on the frequency scale from all the rest. If we are concerned only with the frequencies comparable to the natural frequencies of this cluster of modes, we can ignore all the higher modes and for our purposes we may regard the network as a finite network. At these frequencies the infinitely small meshes into which we could subdivide the individual "inductors" (regions of magnetic energy concentration) and "capacitors" (regions of electric energy concentration) will oscillate in unison in groups.

Briefly we can summarize the above methods of analysis as follows: The medium supporting the electromagnetic field may be regarded as a three-dimensional network of infinitely small meshes in which every mesh is coupled only to the adjacent mesh. Circuit equations applied to this network lead to Maxwell's differential equations. In contrast with this "*natural network model of the medium*" we can construct a *reduced network model* in which only the conductors of the medium are subdivided into meshes. The medium surrounding the conductors is concealed in the mutual impedances of the constituent meshes. Every mesh is coupled to every other mesh and the mutual impedance (or the coupling factor) is

determined from the law of force exerted by a moving charge on a stationary charge. This approach leads to one or two integral equations which can be approximated by a system of linear algebraic equations. While the latter may seem much simpler than the differential equations obtained from the natural network model, in reality their solution would often constitute a much more difficult analytical problem. The natural network model in which each mesh is coupled only to the adjacent meshes is in harmony with the idea of continuous propagation of electromagnetic disturbances; while the reduced network model conforms to the action at a distance philosophy. The difference is merely in the language and ideas and not in substance.

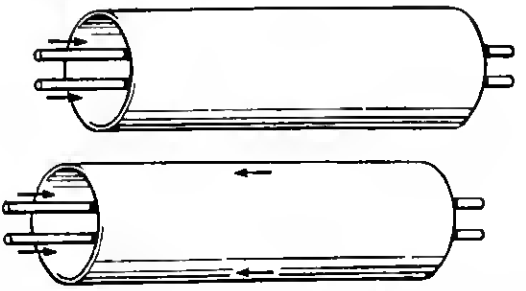

Fig. 4—Two possible modes of propagation in a symmetrically shielded parallel pair.*

Finally, the third method is based on the idea that at certain frequencies, called the natural frequencies, various parts of a closed system oscillate in phase or 180° out of phase, that the most general natural oscillation is the sum of such oscillations, and that the most general forced oscillation can be expressed in terms of fields associated with the natural modes of oscillation. We may call this the *normalized network model of the electromagnetic field*. Thus far we have described it with reference to closed systems or cavity resonators. In effect we have assumed that the amounts of magnetic and electric energy are finite or else we could not talk about T and U functions. The method can be extended to open systems of wave guides.

Modes of Transmission

Let us begin with a coaxial transmission line. Everyone is familiar with the particular mode of transmission in which equal and opposite currents flow in the two conductors. The circuit is completed through the dielectric where the displacement current flows from one conductor to the other. Next, consider a shielded parallel pair. If the structure is symmetric, we shall recognize at once two modes of transmission, Fig. 4. In one mode, the balanced mode, the currents in the wires are equal and opposite; there are

* In the upper part of this figure one of the directional arrows should be reversed.

also equal and opposite currents in the shield which, however, are not equal to the corresponding currents in the wires. In the other mode, the currents in the wires are equal and similarly directed, the return path being through the shield; this mode is similar to the coaxial mode since the wires act in parallel, effectively as one conductor. In the case of n wires there are n distinct modes of transmission. Each mode is characterized by the ratio of currents in the wires and by the field pattern that goes with it.

In all these modes the longitudinal current paths are conductive; but there is no reason whatsoever why the circuit closure should not take place through the dielectric. Even in those modes of transmission in which all longitudinal current paths are conductive, we have to depend on the dielectric for completion of the circuit; this should prepare us for the idea that conductors are not essential for wave transmission. If we include the dielectric, the number of possible longitudinal tubes of flow becomes infinite and so does the number of possible transmission modes; but as the cross-section of each individual tube decreases the longitudinal capacitance also decreases, and these modes will participate in the transfer of power over substantial distances only at correspondingly higher frequencies. It is not merely that at low frequencies the longitudinal impedance becomes very high; it is capacitive and causes high attenuation. The effect is analogous to the attenuation in high-pass filters below the cutoff.

The mathematical analysis which lends quantitative substance to these ideas is similar to that involved in the cavity resonator problem. Once all the modes of transmission have been found, the next problem is that of the excitation of these modes by a given source, that is, of coupling of the source to various modes.

To summarize: A physical transmission line or a wave guide has always an infinite number of transmission modes either independent or substantially independent of each other. It is as if we had a system of single-mode transmission lines without couplers. For each transmission mode the structure behaves as a high-pass filter. If n is the number of conductors, there are $n - 1$ transmission modes with the cutoff frequency equal to zero. Since the lowest non-zero cutoff frequency corresponds to a wavelength comparable to the transverse dimensions of the guide, it is clear that in systems with two or more conductors we have a certain finite number of transmission modes which are well separated on the frequency scale from all the rest. For this reason we may ignore all the higher modes when we are concerned with transmission of low frequencies only, by "low" meaning the frequencies well below the frequency equal to the velocity of light divided by the largest transverse dimension of the transmission line.

Analysis of waves in free space proceeds along similar lines. An electric

"dipole" is the source of the simplest spherical electromagnetic wave. We may picture this dipole as a pair of small spheres connected by a thin rod. Under the influence of an impressed force the charge is made to surge back and forth between the spheres. We cannot have a simple source like a uniformly expanding and contracting sphere as in the case of sound waves. The electric charge is conserved, and the only way we can alter the charge in one place is to transfer it to some other place. A more symmetrical dipole would be a single sphere on the surface of which the charge is made to move back and forth between two hemispheres. Let us call these hemispheres respectively the "northern" and the "southern". When the positive charge accumulates on the northern hemisphere, the radial displacement current flows outwards from it. At the same time an equal radial displacement current flows toward the southern hemisphere. The situation is analogous to the balanced mode of transmission along parallel wires, with the two half spaces acting as "the wires". The distance along the line is the distance from the dipole. The radial transmission line is capacitively loaded but the series capacitance increases as the square of the radius and therefore the capacitive series admittance decreases as the reciprocal of the square of the radius. Hence, at some distance from the dipole, the wave propagation will be quite unimpeded just as in ordinary transmission lines free from loading. Near the dipole the series capacitance is high, and the power carried by the wave in comparison with the energy stored is small.

In the next spherical mode of transmission the polar regions of the spherical generator are similarly charged while the opposite charge is concentrated in the equatorial zone. The zonal character of the radial current distribution persists at all distances from the generator. As might be expected the reactive field in the vicinity of a small "tripole" generator is even stronger than in the case of the dipole source.

The sequence of zonal modes of transmission can be continued indefinitely. Next we could imagine tubular modes in which the space surrounding the generator is subdivided into conical tubes with the radial current in adjacent tubes flowing in opposite directions. This picture is essentially physical; but it corresponds very closely to the mathematical expansion of the general solution of Maxwell's equations in spherical harmonics.

Field Representation in Terms of Fields of Special Types

From the mathematical point of view the method which we have just been considering is based on the idea of representation of the general field in terms of particular fields having certain relatively simple properties. The method is analogous to that employed in circuit theory when the

response to the general electromotive force is expressed in terms of responses to the unit step function, or the unit impulse function, or the steady state responses at various frequencies.

There are numerous variations of the same general idea, some of which are more suitable to one class of problems and others to another class. If the distribution of electric charge and current is known, then in many cases (but not in all) it is best to subdivide it into small volume elements. Except for a possible static electric charge distribution, the elements will be dipoles. The entire field can thus be regarded as the resultant of spherical waves generated by dipoles of given moment and position. To simplify the integration involved in this method certain auxiliary functions, called the retarded potentials, are introduced. One should not try to ascribe to these auxiliary mathematical functions any physical significance and one should always remember that on certain occasions potential functions, other than the retarded potentials, turn out to be more useful. We should also keep in mind that, in order to apply this method, we have to know the complete distribution of electric conduction currents and as a general rule we do not have this information. Consider, for instance, the problem of electromagnetic shielding. The current in the coil is given; but that in the shield has to be determined. There are methods for calculating the induced current; but these methods give at the same time the shielding effectiveness, and that without employing retarded potentials. It is in approximate studies of radiation patterns of antennas and antenna arrays that the retarded potential method is displayed to the best advantage.

The retarded potentials are based on representation of fields in terms of spherical coordinates; that is, in terms of fields associated with hypothetical *point sources* at the origin of the coordinate system. General fields can also be expressed in terms of cylindrical coordinates and, consequently, in terms of fields associated with hypothetical *line* sources situated along the axis of the coordinate system. Likewise, fields can be expressed in cartesian coordinates; that is, in terms of "plane waves". All such representations have useful applications. The current in the coil is given.

Discontinuities

In the analysis of the various transmission modes for a given wave guide it is assumed at first that the boundaries of the wave guide are analytic functions of the coordinates. Any discontinuity or irregularity has to be treated separately, simply because there is nothing in the analytic part of the wave guide to suggest that a discontinuity might occur, or to prescribe the properties of this discontinuity. Discontinuities may be accidental, unavoidable or intentional. A kink in a wire is an example of an accidental

discontinuity. Open air wire lines have to be supported on poles which, together with the insulators, constitute unavoidable discontinuities. The beginning and the end of a line are always present. Usually these latter discontinuities are simply unavoidable; but, in radio, at least one discontinuity, the antenna, is made to serve a useful purpose. It is clear that the generator and the load connected by a two-wire line, Fig. 5, are dipoles which will generate spherical waves as well as the wave guided by the transmission line. At low frequencies the length of the dipoles is so small compared with the wavelength that the field does not reach out into

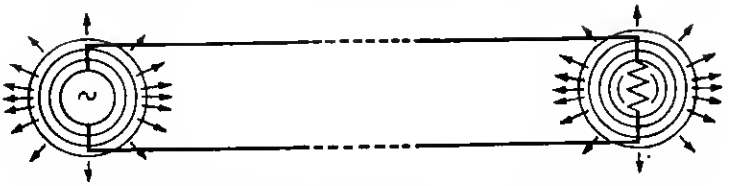

Fig. 5—Formation of spherical waves at the ends of a long pair of parallel wires.

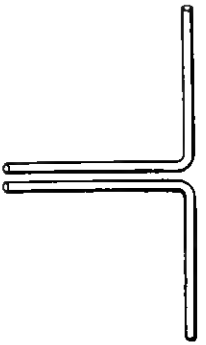

Fig. 6—An antenna.

the region where the radial capacitance becomes negligible and where the spherical wave starts carrying off all the energy that gets there. Spherical waves generated at the beginning and the end of the transmission line are practically stationary waves and constitute merely local reactive reservoirs of energy. The energy is withdrawn from the generator or the transmission line during one half of the cycle only to be returned during the other half. At low frequencies the energy thus exchanged back and forth is so small that normally we don't even think about it. The antenna, Fig. 6, is designed to be a more efficient transformer of the plane wave guided by the parallel pair into the spherical wave which will carry off power to distant points.

Quite frequently discontinuities are introduced intentionally in order to

discriminate against some frequencies. A capacitor in parallel with the wave guide or an inductor in series with it will favor transmission of low frequencies at the expense of high frequencies. These discontinuities are deliberately designed to be sufficiently large to produce noticeable effect. A frequency filter is a more elaborate structure made up of capacitors and inductors designed to achieve desired frequency discrimination.

Discontinuities in high-frequency wave guides are also either accidental, unavoidable or intentional. The principal difference is in the order of magnitude—any irregularity of apparently small physical dimensions may represent a large virtual reservoir of energy. Among the simplest types of intentional discontinuities in wave guides are "irises", Fig. 7. Local fields are created in the vicinity of the irises. Under the influence of a wave traveling along the guide, electric charge and current are induced in the metal partition. On either side of the partition the complete field is the result of the superposition of fields representing various transmission modes. The cutoff frequencies of these modes may be arranged in an

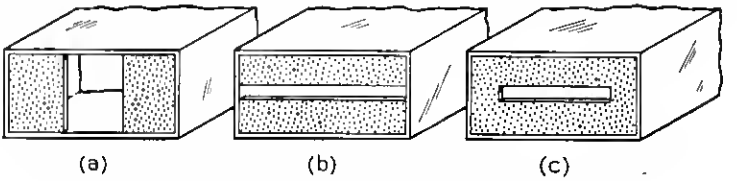

Fig. 7—Inductive, capacitive, and resonant irises.

increasing sequence. If the operating frequency is between the lowest cutoff frequency and the next higher, the propagation constants of all modes except the dominant are real and the corresponding fields will not extend very far from the iris. During one-half cycle the local field withdraws energy from the dominant wave—this being the only source of energy —and during the remaining half this energy is returned. The local field acts as a *virtual source* of power—"virtual" since it operates on borrowed power. On account of symmetry the dominant waves generated by this virtual source and traveling in opposite directions will be of equal intensities. The *scattered wave* traveling toward the source of the incident wave is called *the wave reflected from the iris*; on the other side the scattered and incident waves merge into the *transmitted wave*. The storage of energy in the local field depends on the frequency—hence, the frequency selectivity.

In the case shown in Fig. (7a) the flow of current in the partition is unimpeded and there is no tendency for any local concentration of charge in the partition; the local field is largely magnetic and the iris represents an

inductive reactance. Since any variation of the magnetic field with time always creates an electric field, there will be some capacitance in parallel with the inductance. The same idea may be expressed by saying that the inductance of the iris is not quite independent of the frequency. This lack of constancy is not peculiar to ultra-high frequencies; it is true of coils at low frequencies. Likewise, even at very low frequencies the inductance varies with the frequency because of skin effect.

In the iris shown in Fig. (7b) there are alternating charge concentrations on the upper and lower partitions. The local field is largely electric and the iris is capacitive. A feeble magnetic field associated with charging current is unavoidable, of course; this is also true of capacitors at low frequencies but this time the effect is greater. Finally, an iris of the type shown in Fig. (7c) may be designed to behave as an antiresonant circuit.

In that frequency range in which only the dominant wave is an effective carrier of power to great distances, any discontinuity will behave as a reactive T or Π-network—assuming that observations are made at some distance from the iris where the local field is too feeble to count. This could not be otherwise since there are three parameters at our disposal: two reflection coefficients for waves traveling in opposite directions and one transmission coefficient across the discontinuity. The Reciprocity Theorem requires that the transmission coefficients in the two directions be equal. These three parameters determine the ratios of the reactance elements of the equivalent T or Π-network to the characteristic impedance of the guide.

If the operating frequency exceeds the second cutoff frequency, other waves besides the dominant become effective carriers of power and the equivalent network for the iris becomes more complicated. The iris behaves not only as a dissipative impedance to the dominant wave but also as a negative resistance, to one or more higher order waves.

Boundaries

So far we have paid little attention to the boundaries of the electromagnetic field. Strictly speaking, in any actual situation the field always extends to infinity; the only boundaries there are, are the geometric boundaries between media with different electromagnetic properties. This means that we should solve electromagnetic equations for each homogeneous region, or region with analytically varying properties, and then match the solutions at the boundaries. In many cases, however, this procedure would be very complicated and quite unnecessary. In the case of a cylindrical metal tube with a dipole as a source of power the exact solution may be represented as a particularly formidable integral; but experimentally

we would not be able to detect any difference between the "exact" solution and a much simpler approximate solution.

In the case of rectangular tubes we don't even know how to obtain the "exact" solution in any form; but good approximate solutions are exceedingly simple. The word "exact" is in quotation marks because there can be no really exact solutions of actual physical problems. In the first place the properties of materials are not known exactly; the boundaries between media do not exist in the exact sense of the term; and we just don't know the exact laws of nature. All we really want of any solution is to be accurate enough for some particular purpose. And here is where the idea of idealized boundaries helps in the formulation of simplified, clear-cut mathematical problems. The idea lends flesh and blood to idealized mathematical boundary conditions. *Perfect conductors* have long been mentioned in literature as idealizations of good conductors; but other types of boundaries are of much more recent origin. Perfect conductors are *boundaries of zero surface impedance*; they support electric currents of finite strength when the tangential electric intensity is zero. At these boundaries the tangential magnetic intensity is different from zero. The natural counterpart is a *boundary of infinite impedance* at which the tangential magnetic intensity vanishes but the tangential electric intensity does not. The further generalization is a boundary with a given finite surface impedance which is defined as the ratio of two mutually perpendicular tangential components of the electric and magnetic intensity. The boundary may be isotropic, with its *surface impedance* the same in all directions; likewise, the boundary may be aelotropic. The surface impedance is defined as the ratio of the tangential components of E and H. Since it is necessary to adopt a convention regarding "positive directions" of E and H, these are so chosen that a right-handed screw will advance into the boundary if its handle is turned through 90° from the positive direction of E to coincide with the positive direction of H. In accordance with this convention the positive real part of the surface impedance is associated with an average flow of power into the boundary—that is, with a *passive boundary*. An *active boundary* is a boundary with a negative surface resistance; such boundaries may be used to represent idealized generators of electromagnetic waves and to eliminate from explicit consideration the internal mechanisms of these generators.

Field Equations

Thus far I have tried to present the ideas behind the physical and mathematical analysis of electromagnetic transmission phenomena. These are broader than the electromagnetic laws themselves and, with some super-

ficial modifications, would apply to sound waves, for instance. There are two fundamental equations of transmission of an electromagnetic state, expressing Faraday's law of induction of an electromotive force by a magnetic displacement current and Ampère-Maxwell's law of induction of a magnetomotive force by an electric current. In their most general mathematical form the equations are

$$\begin{aligned} \oint E_s\, ds &= -\frac{\partial}{\partial t}\int\int \mu H_n\, dS, \\ \oint H_s\, ds &= \int\int \rho v_n\, dS + \int\int gE_n\, dS + \frac{\partial}{\partial t}\int\int \epsilon E_n\, dS, \end{aligned} \tag{10}$$

where the subscript s indicates components tangential to a closed path of integration and the subscript n designates components normal to any surface bounded by this closed path. Thus on the left we have "sums" of infinitesimal emf's and mmf's as we travel round some closed curve either on the surface of a wire or just in free space, and on the right we have total magnetic and electric currents linked with this curve. According to our present physical conceptions the magnetic current is always a displacement current defined as the time rate of change of magnetic flux or "displacement". Not that there is anything inconceivable about an actual flow of magnetic charge; it is simply that so far there has been no satisfactory evidence of its existence. In the mathematical analysis it has long been a custom to consider magnetic charges of opposite signs as if they existed; but this is merely for convenience.

The electric current, on the other hand, consists of three components: the *convection current* whose density is the product of the electric charge density ρ and the velocity v; the *conduction current* whose density is proportional to the electric intensity (the gE term in the above equation) and the *displacement current* defined as the time rate of change of the electric displacement. Strictly speaking, tbe conduction current is a convection current but of such a kind that it would be extremely awkward to think of it in terms of charged particles and their velocities.

At the same time the statistical result of the irregular movements of these particles can be expressed, for purposes of transmission of an electromagnetic state, as a continuous movement of charge encountering some resistance. There are, of course, such phenomena as resistance noise which are thus automatically excluded from consideration.

In general to these electromagnetic transmission equations we should add the dynamical equations of motion of electric charge; this is essential when dealing with vacuum tubes. But, in considering passive transmission systems, we either omit the convection current altogether, or else assume

that the velocities of the charged particles are specified, and that the forces which they exert on each other are completely neutralized by the forces external to the field, in which case the convection current appears merely as an "impressed current".

Except for the above restrictions, equations (10) form a complete set; but for mathematical convenience two other equations are usually adjoined. These are

$$\begin{aligned} \iint \epsilon E_n \, dS &= q, \\ \iint \mu H_n \, dS &= 0, \end{aligned} \tag{11}$$

where the double integration is extended over a closed surface. The first of these equations states that the total electric displacement through a closed surface is equal to the net enclosed electric charge; the second denies the physical existence of magnetic charge. These equations can be derived from (10) and for this reason are not quite on the same footing with them.

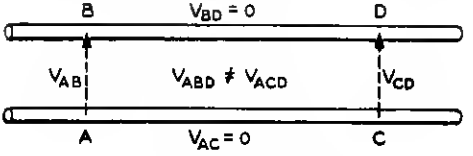

Fig. 8—A pair of parallel wires.

Equation (10) tells us that, except when the field is static, we cannot speak of the electromotive force or the voltage between two points without specifying the path along which we add up the elementary voltages. In fact, equation (10) gives us the difference between the voltages along two different paths connecting the same pair of points. To illustrate, consider a wave along a pair of perfectly conducting wires, Fig. 8. Voltages V_{AC} and V_{BD} along the wires are equal to zero; transverse voltages V_{AB} and V_{CD} are usually unequal; hence $V_{ABD} \neq V_{ACD}$.

If two points are infinitely close, then we can define the voltage unambiguously as the product $E_n ds$ of the electric intensity and the distance between the points. The difference between this voltage and the voltage along any other infinitesimal path is an infinitesimal of the second order, being dependent on the area enclosed by the two paths. In practice two points are sufficiently close if the distance between them is small compared with one quarter wavelength.

Since, except in electrostatics, we cannot speak of the voltage between two points without specifying the path, we cannot speak of the *potential*

difference. In mathematical terms we should say that the differential voltage in a varying electromagnetic field is not an exact differential. To illustrate: $2x\,dx + 2y\,dy$ is an exact differential equal to $d(x^2 + y^2)$ and for this reason its integral depends only on the difference between the values of $(x^2 + y^2)$ at the end points of the path of integration; but $2x\,dx + 2x\,dy$ is not an exact differential and cannot be integrated except when y is given in terms of x so that the path of integration is prescribed.

If equations (10) are applied to infinitesimal closed curves, the following differential equations are obtained:

$$\text{curl } E = -\mu \frac{\partial H}{\partial t}, \qquad \text{curl } H = gE + \epsilon \frac{\partial E}{\partial t}. \tag{12}$$

The expressions curl E and curl H are merely the symbols for the maximum emf's and mmf's per unit area. These equations are not as general as (10) because they assume that E and H are continuous and at least once differentiable. The equations do not hold across the boundary between different media, where they have to be supplemented by the so-called *boundary conditions* which are obtained from (10). Equations (12) do not hold at a wavefront where E and H are discontinuous; there also we have to supplement them by appropriate boundary conditions, which connect the solutions on the two sides of the wavefront.

Analytic Functions

An advance of fundamental importance is made when the field intensities are represented by complex quantities $E\,e^{j\omega t}$ and $H\,e^{j\omega t}$ where ω is the frequency in radians. The equations become

$$\text{curl } E = -j\omega\mu H, \qquad \text{curl } H = (g + j\omega\epsilon)E, \tag{13}$$

and are thus freed from one independent variable, the time t. This does not mean that we have restricted our analysis to steady state fields; Fourier analysis supplies a general rule for passing from steady states to any state whatsoever. Computational difficulties are great but no greater than they would be in any other method.

A still more important advance is made when the field intensities are represented by $E\,e^{pt}$, $H\,e^{pt}$, where the *oscillation constant* $p = \xi + j\omega$ is a complex number. The equations become

$$\text{curl } E = -p\mu H, \qquad \text{curl } H = (g + p\epsilon)E. \tag{14}$$

The solutions of these equations are analytic functions of the complex variable p and a way is open for application of the theory of functions of a complex variable.

Thus if we write

$$E = \sum_{n=0}^{\infty} e_n p^n, \qquad H = \sum_{n=0}^{\infty} h_n p^n, \tag{15}$$

and substitute in (14), we obtain

$$\begin{aligned} &\text{curl } e_0 = 0, \qquad &&\text{curl } e_{n+1} = -\mu h_n, \\ &\text{curl } h_0 = g e_0, \qquad &&\text{curl } h_{n+1} = g e_{n+1} + \epsilon e_n. \end{aligned} \tag{16}$$

If these equations are solved subject to the prescribed boundary conditions, E and H will be expressed as power series in the oscillation constant p.

The function theory has already been used successfully in the *restricted circuit theory*; that is, in the theory of finite networks composed of ideal (independent of the frequency) resistances, inductances and capacitances. Likewise, some very general theorems have been established concerning any *physical* input impedance. Whereas the poles and zeros of a function can be anywhere in the complex p-plane, the poles and zeros of the input impedance of a passive system never lie to the right of the imaginary axis. This leads to a theorem to the effect that all poles and zeros on the imaginary axis are simple. The resistance components of the input impedance on the imaginary axis determine the reactance component and hence the complete impedance function except for a purely reactive impedance. The zeros and poles of an impedance occur always in conjugate pairs. These are some of the general theorems of impedance analysis. Not very long ago I came across an expression for the input impedance of a spherical antenna which was obtained by what appeared superficially as a straightforward conventional method; but as soon as I observed that some poles were situated to the right of the imaginary axis, I knew that the expression had to be false. The existence of poles in this region meant a possibility of oscillations which would increase indefinitely of their own accord.

The difference between finite and infinite networks consists in that the former possess a finite number of zeros and poles. All physical structures always possess an infinite number of such singularities; but a finite number of them may form a cluster in the vicinity of the origin, far removed from all other zeros and poles. When this happens we have a physical finite network. In a reactive network all zeros and poles lie on the imaginary axis. In a slightly dissipative system these zeros and poles move a little to the left of the imaginary axis. This happens, for instance, in the case of a thin antenna. The field in the vicinity of a thin wire is large and the radiated power is only a small fraction of the stored energy. The distribution of poles (the solid circles) and zeros (the hollow circles) is illustrated

in Fig. 9. The zero frequency is always a pole for an open type antenna and a zero for a perfectly conducting loop antenna. As the frequency passes through a zero, the antenna impedance passes through a minimum. As the frequency goes through a pole, the antenna impedance passes through a maximum. The disposition of zeros and poles gives us a qualitative idea of the behavior of the impedance as the frequency varies.

As the radius of the antenna increases, the zeros and poles move farther to the left of the imaginary axis. At the same time some zeros and poles, which for a thin antenna are so far to the left that they have very little effect on the impedance, move nearer the origin. For spherical antennas the number of zeros and poles around the origin is considerably larger than for thin doublets.

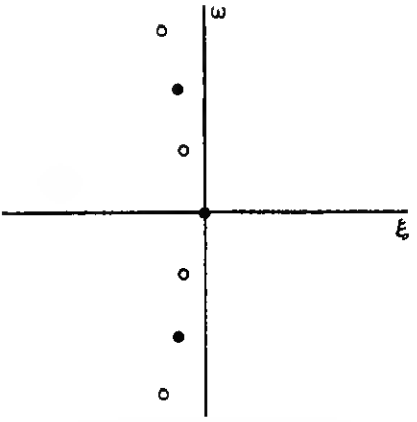

Fig. 9—Distribution of zeros and poles in a dipole antenna: solid circles represent poles; hollow circles zeros.

Circuit and Field Equations

In conclusion I should like to make a few remarks on the relationship between Kirchhoff's circuit equations and Maxwell's field equations. Are the former approximations; and, if so, in what sense? The answer depends on what is meant by Kirchhoff's equations, for their meaning has changed with passing years. It was exactly a hundred years ago that Kirchhoff stated his equations in a kind of postscript to his paper in Poggendorf Annalen; hut he contemplated only the d-c networks. Yet nowadays we interpret these equations in such a way that they are applicable to a-c circuits. Some thirty years went by before Maxwell thus generalized the original Kirchhoff equations with the aid of Lagrange's concepts. Maxwell wrote his circuit equations (not the field equations) in a form applicable only to networks with a finite number of degrees of freedom; but nowadays

we interpret these equations in such a way that we can apply them to one-dimensional transmission lines. In so doing we refrain from making approximations which we normally make when applying Kirchhoff laws to networks of lumped elements. In the latter case it is usual to ignore the inductance of the connecting leads or rather the inductance associated with the loop formed by the leads; but in the case of two-wire transmission lines the "connecting leads" constitute the entire network and the loop inductance is no longer ignored. In the case of lumped networks the capacitance between the connecting leads is normally neglected; but this capacitance is scrupulously included in the analysis of two-wire lines since in this case the "lead capacitance" is all the capacitance there is. And I have already referred to a recent contribution of Kron's who presented a three-dimensional network such that if we apply Kirchhoff's laws to it, we shall obtain Maxwell's field equations. The merger between the two points of view is now complete. In its growth, each theory has developed concepts peculiar to itself. The net result is that we are now in a position to understand electromagnetic phenomena better than ever.